УДК 002.53:004.89

АРХИТЕКТУРА СЕМАНТИЧЕСКОГО WEB-ПОРТАЛА

А.Ф. Тузовский

Институт «Кибернетический центр» ТПУ
Томский научный центр СО РАН
E-mail: TuzovskyAF@kms.cctpu.edu.ru

Предложена архитектура Web-портала на основе использования базовой подсистемы по работе с семантикой информационных объектов. Данная подсистема реализует работу с онтологической моделью и семантическими метаописаниями всех объектов портала, содержащих знания, что позволяет решать задачи по работе с контентом объектов.

Введение

В настоящее время активно развиваются подход по управлению знаниями организаций [1]. Данный подход содержит как организационные методы по активизации создания знаний, по распространению и использованию знаний, так и информационно-программное обеспечение, которое хотя и является не достаточным, но весьма необходимым компонентом. В современных системах управления знаниями множество функциональных компонент объединяется в порталы управления знаниями [2, 3], которые, в качестве основных, включают такие подсистемы, как: база знаний профилей компетентности экспертов и сотрудников компании, поддержка коллективной работы сотрудников и экспертов, формирование структурированного архива документов (библиотеки) и пр. Применение семантических технологий [4] в разработке порталов управления знаниями позволит выполнить их реализацию на качественно новом уровне. Данный подход основывается на использовании единой модели знаний о предметной области организации, согласовании на основе этой модели описаний различных объектов содержащих знания (документов, профилей компетентности и т. п.), поддержке персонализации базы знаний для разных пользователей, повышении точности распространения и поиска требуемых знаний. За счет этого могут быть преодолены многие из существующих барьеров распространения знаний.

Современные Web-порталы

В области информационных технологий порталами называют информационно-программные системы, размещенные в компьютерной сети (Web-приложение) и являющиеся унифицированным средством доступа к набору информационных ресурсов и функций, необходимых определенному сообществу пользователей. Порталы расширяют понятие Web-сайта. Если Web-сайт – это набор логически взаимосвязанных страниц, доступных по протоколу HTTP, то портал – это Web-сайт, который предоставляет широкий набор функций, и в том числе доступ к разнообразным ресурсам.

Основными функциями портала является интеграция информационных источников, приложений и поиск информации. Функциональность портала реализуется на основе некоторой архитектуры, которая определяет согласование функциональных модулей портала, стандартных систем, используемых для их реализации и протоколов взаимодействия. Обобщенная архитектура современных Web-порталов содержит три уровня: интерфейса пользователей, бизнес-логики и данных. На уровне интерфейса пользователей, как правило, используется тонкий клиент (Web-браузер), способный визуализировать графическое представление, описанное на языке HTML. Для выполнения некоторых дополнительных действий могут использоваться и другие приложения (клиент электронной почты, RSS-клиент и т. п.). Для реализации инфраструктуры портала применяется ряд широко известных программных систем и технологий, например, таких, как Microsoft IIS, Apache HTTP Server, Oracle Application Server. В качестве хранилищ данных используются реляционные базы данных, доступ к которым осуществляется с применением технологий OLEDB, ODBC, JDBC и т. п. В качестве технологий программирования могут использоваться такие технологии, как ASP, ASP.NET, PHP, JSP. Визуальное представление описывается на языке HTML, который интерпретируется Интернет-обозревателем пользователя.

Хотя подходы к разработке и реализации порталов можно считать достаточно проработанными с точки зрения методов и используемых технологий, существует объективная необходимость их развития. Эта необходимость обусловлена, с одной стороны, развитием телекоммуникационных технологий, делающих информацию принципиально более доступной, а с другой, лавинообразным ростом объема информации, в которой все сложнее отыскать необходимую информацию в заданном контексте.

Портал является такой информационной системой, которая организует унифицированный доступ к целевому информационному пространству, и поэтому проблема повышения эффективности информационных процессов при больших объемах информации стоит особенно остро. Одним из подходов к решению данной проблемы является переход на семантический уровень при сборе, обработке, накоплении, хранении, поиске и распространении информации.

Архитектура семантического портала

В настоящее время активно ведутся научные исследования по дальнейшему развитию Web-порталов. Одним из основных направлений этих исследований является переход от методов работы с синтаксисом к методам работы с семантикой. Методы работы с семантикой, и их программная реализация, называются семантическими технологиями [4]. Использование семантических технологий рассматривается в качестве средства повышения качества, адаптируемости и функциональности создаваемых порталов [5].

Порталы, которые используют для реализации своих функций семантические технологии, получили название семантических порталов. Семантические технологии используют в создаваемых порталах для реализации различных функций, например, таких как проектирование подсистем портала, организация навигации, взаимодействия пользователей, работы с документами.

В данной статье предлагается подход к созданию семантического портала, в котором семантические технологии используются для реализации набора базовых функций, составляющих семантическую подсистему портала. Архитектура предлагаемого семантического портала показана на рис. 1. Семантическая подсистема разработана с целью описания объектов портала на семантическом уровне (в отличие от синтаксического уровня).

Переход на семантический уровень осуществляется за счет устранения синтаксической многозначности и учета связей между понятиями. Для этого семантическая подсистема в процессе обработки информации предоставляет возможности учета:

1) Наличия эквивалентных лексических конструкций (синонимов) в естественном языке. Это достигается за счет описания понятий в онтологии множественными лексическими конструкциями. В результате появляется возможность сравнения синтаксически различной, но семантически схожей информации.
2) Иерархической природы понятий, выражающейся в виде отношения «класс – подкласс». Это достигается за счет использования таксономии понятий в онтологической модели.

Возможности семантической подсистемы используются для структуризации информации в портале и предоставлении ее пользователям. Предлагаемое семантическая подсистема поддерживает следующие варианты использования онтологии: аннотирование объектов; семантический поиск; формирование списка объектов, связанных с исходным объектом; формирование списка объектов, семантически близкий к исходному объекту.

В качестве моделей представления знаний в данной работе используются онтологии предметных областей [6]. Онтология содержит описания понятий предметной области и отношений между ними. Онтология определяется как знаковая система $O=\{C, R, L, P_c, P_{LC}, P_{LR}\}$, в которой $C=\{c_1,...,c_n\}$ – конечное множество понятий в онтологии; $R=\{r_1,...,r_m\}$ – конечное множество бинарных отношений $r_i(c_x,c_y)$ между понятиями; $L=\{l_1,...,l_k\}$ – конечное множество лексических меток (словарь онтологии); $P_c \subseteq C \times C$, $P_c \in R$ – антисимметричное, транзитивное, нерефлексивное бинарное отношение, являющееся отношением частичного порядка на множестве понятий C, $P_{LC} \subseteq L \times C$ и $P_{LR} \subseteq L \times R$ – бинарные отношения инцидентности между множе-

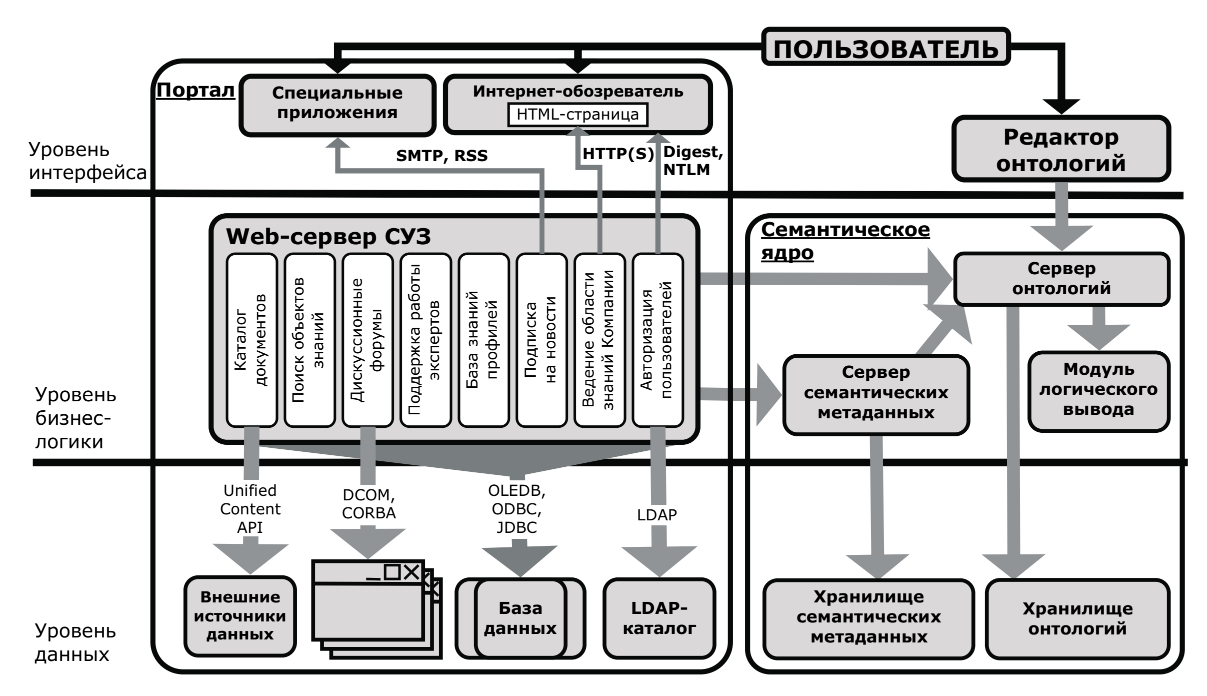


Рис. 1. *Архитектура семантического портала*

ствами L и C, а также L и R. Для записи в портале онтологических моделей используется язык OWL-DL [7].

В портале имеется большое количество информационных объектов $D=\{d_1, \ldots, d_p\}$, таких как документы, сообщения, новости, адреса ресурсов, профили пользователей, описания функциональных сервисов и т. п. Элементы онтологии используются для описания семантики информационных объектов в виде семантических метаданных. Семантические метаданные для объекта d_i представляют собой: $M(d_i)=\{(\mathrm{tr}_{i1},k_{i1}),\ldots,(\mathrm{tr}_{ij},k_{ij})\}$, где $\mathrm{tr}_{ij}=\langle s_{ij},p_{ij},o_{ij}\rangle$ – триплет; $s_{ij}\in C\cup I$ – субъект в триплете; $p_{ij}\in R\cup A$ – предикат в триплете; $o_{ij}\in C\cup I\cup V$ – объект в триплете; $k_{ij}\in(0,1]$ – коэффициент, обозначающий релевантность триплета tr_{ij} объекту d_i.

Для записи в портале онтологических моделей используется язык RDF (*Resource Definition Framework*) [8]. Использование онтологий и семантических метаданных позволяет определить семантическую (смысловую) близость объектов портала. Для этого могут быть использованы методы оценки семантической близости элементов онтологии [1, 9]. Для автоматического определения положения элементов в таксономии используются алгоритмы автоматического вывода в модели дескриптивной логики $\mathrm{SHIQ(D_n)^-}$, т. к. дескриптивная логика является базовым формализмом описания онтологий на языке OWL-DL [10]. На основе оценки семантической близости различных элементов метаданных разработаны методы расчета семантической близости метаданных $SM\,(M(d_i),\, M(d_j))$.

Состав и структура семантической подсистемы

Функциональность семантической подсистемы по работе с онтологиями расположена в сервере онтологий, а функциональность по работе с семантическими метаданными – в сервере семантических метаданных. ***Сервер онтологий*** – это отдельно функционирующее приложение, хранящее множество онтологий и предоставляющее к ним доступ. Сервер онтологий используется различными приложениями, которым нужны различные функции в процессе реализации жизненного цикла онтологии: создания, оценки, использования и последующей актуализации. Для реализации в семантической подсистеме выбранных вариантов использования онтологии, сервер онтологий предоставляет следующие функции: хранение онтологий; извлечение онтологий; логический вывод; поиск в онтологии запрашиваемых понятий и отношений. ***Сервер семантических метаданных*** – это отдельно функционирующее приложение, хранящее семантические метаданные, предоставляющее к ним доступ и обрабатывающее их. Для реализации выбранных вариантов использования онтологии сервер семантических метаданных, тесно взаимодействуя с сервером онтологий, предоставляет следующие функции: составление семантических метаданных; хранение семантических метаданных; извлечение семантических метаданных; сравнение семантических метаданных.

Созданные с помощью функций семантической подсистемы онтологии предметных областей и семантические метаданные объектов семантического портала используются при обработке объектов. При аннотировании устанавливается соответствие элементов информационного описания объектов семантического портала с элементами онтологии. Результатом аннотирования являются семантические метаданные объектов семантического портала. Семантические метаданные являются основой для реализации трех других функций семантической подсистемы.

Возможность определения семантической близости позволило разработать методы решения таких задач семантической подсистемы (рис. 2) важных для работы порталов, как:

- семантический поиск на основе близости семантического запроса $Q=(q_1\wedge q_2\wedge\ldots\wedge q_n)$, где

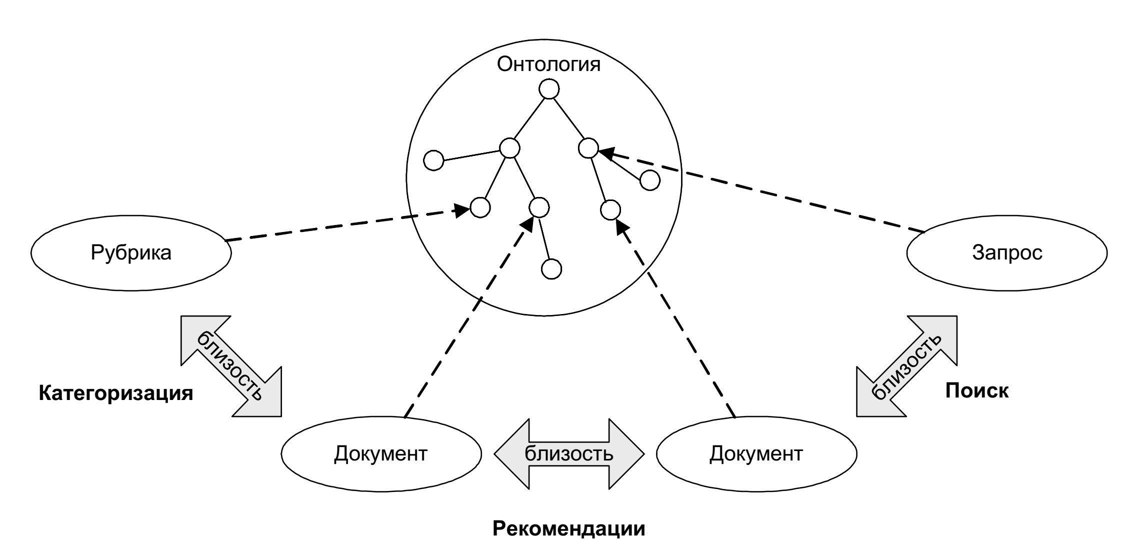


Рис. 2. *Использование алгоритма оценки семантической близости*

$q_i=(s_i, p_i, o_i)$, и семантических метаданных всех информационных объектов портала.

- классификации информационных объектов в соответствии с набором рубрик на основе семантических описаний рубрик библиотеки $H=(h_1 \wedge h_2 \wedge ... \wedge h_n)$, где $h_i=(s_i, p_i, o_i)$ и семантических метаданных всех информационных объектов портала.
- рекомендаций пользователям портала на основе семантических описаний метаописаний профилей пользователей $U=(u_1 \wedge u_2 \wedge ... \wedge u_n)$, где $u_i=(s_i,p_i,o_i)$ и семантических метаданных всех информационных объектов портала.

Для выполнения семантического поиска информационных объектов необходимо наличие семантических метаданных у объектов и представление поискового запроса пользователя с помощью элементов онтологии. Поиск выполняется путем сравнения семантических метаданных с запросом. Объект считается релевантным запросу в том случае, когда в его семантических метаданных содержаться все элементы из запроса или подклассы этих элементов. Таким образом, при поиске учитывается иерархия понятий предметной области. Функция формирования списка объектов, связанных с исходным объектом в семантическом портале может иметь различное применение. В разработанном семантическом портале на ее основе реализована категоризация. Для выполнения категоризации объектов с учетом их семантики необходимо наличие семантических метаданных у объектов и у категорий, к которым нужно отнести объекты. Предполагается, что все множество возможных категорий будет иерархически упорядочено. Категоризация выполняется путем сравнения семантических метаданных объекта и категории. Объект считается относящимся к категории, если в его семантических метаданных содержаться хотя бы некоторые элементы из семантических метаданных категории или подклассы этих элементов. Очевидно, что объект может быть отнесен к одной или более категории.

Функция формирования списка объектов, семантически близких к исходному объекту также может иметь различное применение в семантическом портале. В разработанном семантическом портале данная функция использовалась при реализации рекомендаций. Функция рекомендации заключается в предоставлении пользователю объектов, похожих по содержанию на некоторый объект, фигурирующий в запросе. Выполнение поиска похожих объектов осуществляется на основании сравнения семантических метаданных объектов.

Заключение

Спроектированная семантическая подсистема портала была программно реализована на платформе Microsoft.Net с использованием языка программирования C# и технологии .Net Remoting. В качестве модуля выполнения автоматического вывода в моделях дескриптивной логики, соответствующих онтологиям предметных областей, использована система RACER[11]. С использованием созданного программного обеспечения было выполнено тестирование разработанных методов, которое показало высокий уровень формальной полноты и точности всех алгоритмов поиска, категоризации и рекомендаций.

Разработанное программное обеспечение семантической подсистемы апробировано в процессе разработки и внедрения двух семантических Web-порталов различного уровня: портала «Petroleum Engineers Virtual Network» совместного российско-шотландского подразделения Томского политехнического университета и портала поддержки корпоративной системы управления знаниями компании «ЭлеСи» (г. Томск).

СПИСОК ЛИТЕРАТУРЫ

1. Тузовский А.Ф., Чириков С.В., Ямпольский В.З. Системы управления знаниями (методы и технологии). – Томск: Изд-во НТЛ, 2005. – 260 с.
2. Collins H. Enterprise knowledge portals: next generation portal solutions for dynamic information access, better decision making and maximum results. – N.Y.: AMACOM, 2003. – 430 p.
3. Firestone J.M. Enterprise information portals and knowledge management. – Oxford: Butterworth-Heinemann, 2003. – 422 p.
4. TopQuadrant Technology briefing. Semantic technology [Электронный ресурс]. – 2004. – Режим доступа: http://www.topquadrant.com/documents/TQ04_Semantic_Technology_Briefing.PDF
5. Lausen H., Stollberg M. Semantic Web Portals – State of the Art Survey technology [Электронный ресурс]. – 2004. – Режим доступа: http://www.deri.at/publications/techpapers/documents/DERI-TR-2004-04-03.pdf
6. Guarino N. Understanding, building and using ontologies // International Journal of Human-Computer Studies archive. – 1997. – V. 46. – № 2–3. – P. 293–310.
7. Web Ontology Language. Overview // [Электронный ресурс]. – 2003. – Режим доступа: http://www.w3.org/TR/owl-features/
8. W3C, «RDF/XML Syntax Specification (Revised)» // [Электронный ресурс]. – 2003. – Режим доступа: http://www.w3.org/TR/rdf-syntax-grammar/
9. Тузовский А.Ф., Васильев И.А., Усов М.В. Программная реализация основных компонент информационно-программного обеспечения системы управления знаниями // Известия Томского политехнического университета. – 2004. – Т. 307. – № 7. – С. 116–122.
10. The Description Logic handbook: theory, implementation, applications / Ed. F. Baader. – Cambridge: Cambridge University Press, 2003. – 564 p.
11. Haarslev V., Moller R. RACER: A core inference engine for the Semantic Web // Proc. of the 2nd Intern. workshop on evaluation of ontology-based tools (EON-2003). – Florida, USA, 2003. – P. 27–36.